10231 Abstracts Collection
Structure Discovery in Biology: Motifs, Networks & Phylogenies
— Dagstuhl Seminar —

Alberto Apostolico[1], Andreas Dress[2] and Laxmi Parida[3]

[1] Georgia Institute of Technology, US
axa@cc.gatech.edu
[2] Shanghai Institutes for Biological Sciences, CN
andreas@picb.ac.cn
[3] IBM TJ Watson Research Center, US
parida@us.ibm.com

Abstract. From June 6 to June 11, 2010, the Dagstuhl Seminar 10231 "Structure Discovery in Biology: Motifs, Networks & Phylogenies " was held in Schloss Dagstuhl – Leibniz Center for Informatics. During the seminar, several participants presented their current research, and ongoing work and open problems were discussed. Abstracts of the presentations given during the seminar as well as abstracts of seminar results and ideas are put together in this paper. The first section describes the seminar topics and goals in general. Links to extended abstracts or full papers are provided, if available.

Keywords. Computational biology, form, structure, function, combinatorics, algorithms

10231 Executive Summary – Structure Discovery in Biology: Motifs, Networks & Phylogenies

This seminar was intended to focus on combinatorial and algorithmic techniques of structure discovery from biological data that is at the core of understanding a coherent body of data, small or large. In biological systems, similarly to the tenet of modern architecture, "form and function are solidly intertwined". Thus, to gain complete understanding in various contexts, the curation and study of forms turns out to be a mandatory first phase.

Biology is in the era of the Ome's: Genome, Proteome, Transcriptome, Metabolome, Interactome, ORFeome, Recombinome and so on. Thus it is no surprise that biological data is accumulating at a much faster rate than it can be understood. While on one hand, the sheer size of data can be daunting, this provides on the other hand a golden opportunity for testing (bioinformatic) structure-discovery primitives and methods. In spite of the difficulties of structure discovery, there are reasons to believe that evolution based on reproduction, variation,

and selection endowed biological systems with some underlying principles of organization (involving redundancy, similarity, and so on) that appear to be present across the board. Correspondingly, it should be possible to identify a number of primitive characteristics of the various embodiments of form and structure (using for instance notions of "maximality", "irreducibility", etc.) and build similarly unified discovery tools around them. At a core level in these efforts, there is the need for novel techniques enabling the automated discovery of structures, whether syntactic, such as patterns or motifs, or semantic, such as phylogenies. It is therefore a worthwhile effort to try and identify these primitives.

The workshop gathered a roster of highly qualified senior participants and several talented young researchers who convened to discuss issues of modeling, formalization, and algorithmic design as they emerge in the discovery of biological structure. The schedule unraveled in form of a busy sequence of sessions, intermixed with small group-work sessions and guest lectures. The opening consisted of an extended round-robin aimed at letting participants give detailed accounts of their background and interests. Following opening remarks by Andreas Dress and completed by concluding comments by Alberto Apostolico, three broad-spectred lectures were delivered by Benny Chor, Laxmi Parida, and Raffaele Giancarlo in the afternoon of the first day. A special lecture on epigenetics offered on Tuesday morning in the concurrent Workshop on "The Semantic of Information" provided the opportunity to nicely complement the presentations already scheduled. The morning session resumed with three lectures centered on networks, offered in turn by Axel Mosig, Ina Koch, and David Gilbert. The afternoon session was devoted to phylogenies, with featured presentations by Stefan Gruenewald, David Bryant, Peter F. Stadler, and Andeas Spillner. After dinner, a guest lecture by Michael Clausen exposed, not without entertaining notes, the analogies between motifs in biosequences and music and methods for searching for them. The morning session of Wednesday revolved on regulation, with presentations by Rahul Siddharthan, Esko Ukkonen, Matteo Comin, and Ion Mandoiu. Thursday was entirely dedicated to motif discovery in various contexts. Presentations were given in the morning by Jens Stoye, Peter Erdos, Asif Javed, and Nadia Pisanti, followed in the afternoon by Cinzia Pizzi, Fabio Cunial, Matthias Gallé and Benjarath Pupacdi. After dinner, a thorough introduction to toponomics by Andreas Dress was followed by a lecture and software demonstration by Peter Serocka. The participants re-convened in the lecture hall on Friday morning for a general discussion, an assessment of the experience and recommendation for (enthusiastically endorsed) possible encores in the future.

Keywords: Computational biology, form, structure, function, combinatorics, algorithms

Joint work of: Apostolico, Alberto; Dress, Andreas; Parida, Laxmi

The tight span of a diversity

David Bryant (University of Auckland, NZ)

The tight span is a simple and elegant construction that takes a metric and returns a cell complex (or a split network). Andreas Dress's epic "Trees, Tight Extensions of Metric Spaces, and the Cohomological Dimension of Certain Groups: A Note on Combinatorial Properties of Metric Spaces" should rightly viewed as the ancestor of all split based phylogenetic networks (spectronet, NeighborNet, Q-Net, split decompositon etc.). The construction has many fascinating and elegant properties, but the attraction for us phylogeneticist mathematicians is that when a metric is treelike, the method returns a tree, and as the metrics get less tree-like, the method returns networks that look less and less treelike.

Nevertheless, distance based methods have their limitations. For several years we've been looking for ways of defining tight spans on pattern distributions, a statistically responsible tight span. Very very recently, we made a major step forward. We have found what we think is the appropriate tight span definition for diversity functions, generalisations of metrics that assign diversity values (like phylogenetic diversity) to subsets of the taxa set. I introduce the tight span for metrics, then give our definition for diversity measures, talk about its tight span and hint how this could provide a way to consistently construct networks under a general Markov model.

Keywords: Tight span, phylogenetic networks, diversity functions

Joint work of: Bryant, David; Tupper, Paul F. (Dept. of Mathematics, Simon Fraser University)

Approaches to Large Whole Genome Pylogenies, and Initial Results

Benny Chor (Tel Aviv University, IL)

Short genomic sequences (e.g. genes or proteins) have driven early bioinformatics research. With the advent of various high throughput biotechnologies, sequences have taken a back seat. But with hundreds of complete genome sequences known (and thousands in the making), we now face new algorithmic challenges, and are able to approach questions that were inaccessible before. I describe a number of questions and results in this field, concentrating on approaches to construct whole genome phylogenies from large eukaryotic genome sequences. Some disturbing preliminary initial results are sketched, which highlight the need for alternative sequence measures capturing both short range edit distance like measure and longer range genome rearrangement like measures.

Keywords: Whole genome phylogenies, eukaryotic genomes

Joint work of: Chor, Benny; Burstein, David; Tuller, Tamir; Ulitsky, Igor; Cohen, Eual; Pasmanik-Chor, Metsada (Tel Aviv University)

Remote Homology Detection of Protein Sequences

Matteo Comin (University of Padova, IT)

The classification of protein sequences using string kernels provides valuable insights for protein function prediction. Almost all string kernels are based on patterns that are not independent, and therefore the associated scores are obtained using a set of redundant features. In this talk we will discuss how a class of patterns, called Irredundant, is specifically designed to address this issue. Loosely speaking the set of Irredundant patterns is the smallest class of independent patterns that can describe all patterns in a string. We present a classification method based on the statistics of these patterns, named Irredundant Class. Results on benchmark data show that Irredundant Class outperforms most of the string kernel methods previously proposed, and it achieves results as good as the current state-of-the-art methods with a fewer number of patterns. Unfortunately we show that the information carried by the irredundant patterns can not be easily interpreted, thus alternative notions are needed.

Keywords: Classification of protein sequences, irredundant patterns

Joint work of: Comin, Matteo; Verzotto, Davide (Department of Information Engineering, University of Padova, Italy)

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2010/2741

The subsequence composition of polypeptides

Fabio Cunial (Georgia Institute of Technology, US)

The quantitative underpinning of the information content of biosequences represents an elusive goal and yet also an obvious prerequisite to the quantitative modeling and study of biological function and evolution. Several past studies have addressed the question of what distinguishes biosequences from random strings, the latter being clearly unpalatable to the living cell. Such studies typically analyze the organization of biosequences in terms of their constituent characters or substrings and have, in particular, consistently exposed a tenacious lack of compressibility on behalf of biosequences. This research attempts, perhaps for the first time, an assessement of the structure and randomness of polypeptides in terms on newly introduced parameters that relate to the vocabulary of their (suitably constrained) *subsequences* rather than their substrings. It is shown that such parameters grasp structural/functional information, and are related to each other under a specific set of rules that span biochemically diverse polypeptides. Measures on subsequences separate few amino acid strings from their random permutations, but show that the random permutations of most polypeptides amass along specific linear loci.

Keywords: Information content of polypeptides, constrained subsequences, suffix graph

Joint work of: Cunial, Fabio; Apostolico, Alberto

See also: A. Apostolico, F. Cunial. "The subsequence composition of polypeptides". Journal of Computational Biology, 2010.

Modelling complex dynamical systems with cellular automata

Andreas Dress (Shanghai Institutes for Biological Sciences, CN)

In recent years, Lin Wei, Peter Serocka, and Andreas Dress began re-investigating structure-formation processes in coupled oscillating 2D systems based on the Ideal-Storage Model introduced in 1982 (see Andreas W.M. Dress and LIN Wei: Dynamics of a discrete-time model of an "ideal-storage" system describing hetero-catalytic processes on metal surfaces, submitted). Our analysis showed that, from the dynamical-systems point of view, these systems provide an amazingly rich class of toy examples for studying the onset of oscillating and chaotic behaviour.

As was observed in the 1980's by Heike Schuster and Martin Gerhardt in their PhD theses dealing with heterogeneous catalytic processes on metal surfaces, this becomes even more apparent when the dynamic behaviour of a coupled rectangular array of many such systems is studied, placed e.g., in the fashion of a cellular automaton, at the squares of a rectangular grid and coupled by diffusion relative to one of its two parameters.

The resulting patterns were published by Scientific American in 1988 in its "Computer-Recreation" section and, a little later, similar patterns were actually observed experimentally by Ronald Imbihl (currently Institut fur Physikalische Chemie und Elektrochemie, Leibniz Universitaet Hannover), then working at Gerhard Ertl's lab in Berlin who received the Nobel prize in 2007 for his work on chemical processes on solid surfaces.

Furthermore, it turned out in subsequent studies that one can use such CAs for modeling all sorts of complex processes, from phase transition in binary mixtures (see e.g. Vannozzi, C., Fiorentino, D., D'Amore, M., Rumschitzki, D.S., Dress, A.W.M., Mauri, R. (2006): Cellular automata model of phase transition in binary mixtures. Industrial & Engineering Chemistry Research, 45 (8), 2892-2896) to using them as a "metaphor" for cancer onset caused by only one short pulse of "tissue dis-organization" (changing e.g. for only one single time step the diffusion coefficient) as hypothesized by Ao Ping in his recent papers questioning the current gene/genome centric view on cancer onset.

Keywords: Complex dynamical systems, cellular automata, structure formation

Joint work of: Dress, Andreas; Serocka, Peter; Hordijk, Wim; Wei, Lin

The Ideal Storage Cellular Automaton Model

Andreas Dress (Shanghai Institutes for Biological Sciences, CN)

We have implemented and investigated a spatial extension of the original ideal storage model by embedding it in a 2D cellular automaton with a diffusion-like coupling between neighboring cells. The resulting ideal storage cellular automaton model (ISCAM) generates many interesting spatio-temporal patterns, in particular spiral waves that grow and compete" with each other. We study this dynamical behavior both mathematically and computationally, and compare it with similar patterns observed in actual chemical processes. Remarkably, it turned out that one can use such CA for modeling all sorts of complex processes, from phase transition in binary mixtures to using them as a metaphor for cancer onset caused by only one short pulse of 'tissue dis-organzation' (changing e.g. for only one single time step the diffusion coefficient) as hypothesized in recent papers questioning the current gene/genome centric view on cancer onset by AO Ping et al.

Joint work of: Dress, Andreas; Serocka, Peter; Hordijk, Wim; Wei, Lin

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2010/2728

Balanced Vertices in Trees and a Simpler Algorithm to Compute the Genomic Distance

Peter Erdos (Alfréd Rényi Inst. of Mathematics - Budapest, HU)

This paper provides a short and transparent solution for the covering cost of white–grey trees which play a crucial role in the algorithm of Bergeron *et al.* to compute the rearrangement distance between two multichromosomal genomes in linear time (*Theor. Comput. Sci.*, 410:5300–5316, 2009). In the process it introduces a new *center* notion for trees, which seems to be interesting on its own.

Keywords: Genomic distance, rearrengment, balanced points in trees

Joint work of: Erdos, Peter; Soukup, Lajos (Alfréd Rényi Institute for Mathematics, Budapest, Hungary); Stoye, Jens (Universitat Bielefeld, Technische Fakultat, AG Genominformatik, Bielefeld, Germany)

Full Paper:
http://arxiv.org/abs/1004.2735v1

See also: arXiv 1004.2735v1 [cs:DM]

A New Tree Distance Metric for Structural Comparison of Sequences

Matthias Gallé (INRIA - Rennes, FR)

In this paper we consider structural comparison of sequences, that is, to compare sequences not by their content but by their structure. We focus on the case where this structure can be defined by a tree and propose a new tree distance metric that capture structural similarity. This metric satisfies non-negativity, identity, symmetry and the triangle inequality. We give algorithms to compute this metric and validate it by using it as a distance function for a clustering process of slightly modified copies of trees, outperforming an existing measure.

Keywords: Tree distance, structure discovery, Parseval metric, Tanimoto distance

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2010/2737

Functional Information, Biomolecular Messages and Complexity of BioSequences and Structures

Raffaele Giancarlo (Universitá di Palermo, IT)

In the quest for a mathematical measure able to capture and shed light on the dual notions of information and complexity in biosequences, Hazen et al. have introduced the notion of Functional Information (FI for short). It is also the result of earlier considerations and findings by Szostak and Carothers et al. Based on the experiments by Charoters et al., regarding FI in RNA binding activities, we decided to study the relation existing between FI and classic measures of complexity applied on protein-DNA interactions on a genome-wide scale. Using classic complexity measures, i.e, Shannon entropy and Kolmogorov Complexity as both estimated by data compression, we found that FI applied to protein-DNA interactions is genuinely different from them. Such a fact, together with the non-triviality of the biological function considered, contributes to the establishment of FI as a novel and useful measure of biocomplexity. Remarkably, we also found a relationship, on a genome-wide scale, between the redundancy of a genomic region and its ability to interact with a protein. This latter finding justifies even more some principles for the design of motif discovery algorithms. Finally, our experiments bring to light methodological limitations of Linguistic Complexity measures, i.e., a class of measures that is a function of the vocabulary richness of a sequence. Indeed, due to the technology and associated statistical preprocessing procedures used to conduct our studies, i.e., genome-wide ChIP-chip experiments, that class of measures cannot give any statistically significant indication about complexity and function. A serious limitation due to the widespread use of the technology.

References

1. J.M. Carothers, S.C. Oestreich, J.H. Davis, and J.W. Szostack. Informational complexity and functional activity of RNA structures. J. AM. CHEM. SOC., 126 (2004), pp. 5130-5137.
2. R.M. Hazen, P.L. Griffin, J.M. Carothers, and J.W. Szostak. Functional Information and the emergence of biocomplexity. Proc. of Nat. Acad. Sci, 104 (2007), pp. 8574-8581.
3. J.W. Szostak. Functional Information: molecular messages, Nature, 423 (2003).

Keywords: Functional activity, sequence complexity, combinatorics on words, protein-DNA interaction

Joint work of: Giancarlo, Raffaele; Corona, Davide (Dulbecco Telethon Institute c/o Università di Palermo, Dipartimento di Biologia Cellulare e dello Sviluppo, Palermo, Italy); Di Benedetto, Valeria; Gabriele, Alessandra; Utro, Filippo (Dipartimento di Matematica e Informatica, Palermo, Università di Palermo, Italy)

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2010/2688

Patterns, Systems and Synthetic Biology

David Gilbert (Brunel University, GB)

I give a brief overview of the classes of data that can be used as bases for pattern construction in Systems Biology, namely: (1) sequence: DNA, RNA; (2) biochemical structure: DNA, RNA, proteins; (3) interactions: protein-protein, etc. (4) behaviour: gene regulation, metabolic networks, signal transduction; (5) phenotypes: cell, tissue, organ, etc. often obtained by imaging; (6) multiscale patterns of data from systems at several levels. I then discuss the use of patterns: (1) characterisation – for understanding? (2) classification – searching for new instances? (3) biomodel engineering – to enable the construction of models of biological systems; (4) synthetic biology – as a guide construction of (bio)systems? I illustrate this with examples from the work in my group, with a focus in the area of signal transduction pathways.

Keywords: Patterns, biomodel engineering, systems biology, synthetic biology, signal transduction networks, metabolic pathways, multiscale modelling

Joint work of: Gilbert, David (Brunel University); Gao, Qian (Brunel University); Tree, David (Brunel University); Trybilo, Maciej (Brunel University); Harvey, Amanda (Brunel University); Wu, Zujian (Brunel University); Heiner, Monika (Cottbus University); Donaldson, Robin (Glasgow University); Breitling, Rainer (Glasgow University)

A Quartet Version of Split Decomposition

Stefan Gruenewald (Shanghai Institutes for Biological Sciences, CN)

Split decomposition is one of the first and most widely used methods to reconstruct not necessarily compatible weighted split systems which can be visualized as undirected phylogenetic networks (split netwoks). Its input is a dissimilarity function (usually a metric) but it can also be considered as a quartet-based method where, for all 4 taxa, the weight of at least one of the three possible quartets is zero. I introduce some work in progress where we develop a variant of split decomposition such that the weights of all quartets (which might be computed directly from the raw data) can be positive. The method can reconstruct more general than weakly compatible split systems. Further, a systematic bias towards too long quartets can be recognized from the output while too long distances will cause long pending edges in the split decomposition network. Some first experiments show that the output of the new method depends heavily on how the quartet weights are computed, and it tends to be significantly different from the (distance-based) split decomposition and other methods like NeighborNet and QNet.

Keywords: Phylogenetic trees, phylogenetic networks, quartets, splits, split decomposition

Joint work of: Gruenewald, Stefan; Ma, Ningning; Yang, Jialiang

Recombinomics

Asif Javed (IBM TJ Watson Research Center, US)

Genetic recombinations plays a vital role in the physical health of an individual as well as the collective evolutionary health of the species. Despite their undeniable importance, the recombinational dynamics of the genome are poorly understood. Recombinations bring together sequences with divergent pasts and the computational task of untangling potential phylogenies is very challenging. Most phylogenetic studies thus avoid it by relying on the non-recombining loci: Y chromosome and mitochondrial DNA. But these loci comprise less than 2% of the genome and hence cannot paint the complete phylogenetic landscape. We aim to identify past recombinations from their footprints in extant sequences. Our results indicate that the unique junctions, created by crossovers among dissimilar sequences, bear witness to these past events. These historic recombinations can be detected from the current sequences, and carry evidence of the mutual past of the sequences.

Keywords: Recombinations, ancestral recombination graphs

Joint work of: Javed, Asif; Melé, Marta; Pybus, Marc; Calafell, Francesc; Bertranpetit, Jaume; Parida, Laxmi

Concepts for Modularization of Biochemical Networks based on Transition-Invariant Analysis

Ina Koch (Universität Frankfurt, DE)

Background: Automatic modularization with biological meaning is still a challenging problem in systems biology. Modules can be derived at different levels of abstraction, at sequence level, structural level, or network level. We present concepts which are based on T-invariant analysis and can be applied to metabolic as well as to signal transduction and gene regulatory networks.

Although of the new high-throughput methods, in many cases there are not enough quantitative data available for the biochemical system of interest. Because of the lack of kinetic data discrete approaches have been developed and applied for modeling. Among these methods, Petri net formalism has been applied for modeling and solving many biological questions [1, 2, 3, 4]. In the context of Petri net formalism, besides an intuitive visualization, animation techniques and analysis methods several modularization techniques can be used, we want to discuss here.

Methods: Petri nets are bipartite directed graphs which consist of two types of vertices, places (chemical compounds, for example, metabolites) describing the passive sites and transitions representing active sites (chemical reactions, interactions) of the network. T-invariants are based on the incidence matrix of Petri nets. Based on the incidence matrix two linear equation systems can be derived, where the solutions correspond to vectors of places or vectors of transitions, respectively. To represent the steady state assumption we set all equations to zero. Now, we are interested in nontrivial, semi-positive, and minimal solutions, called just T-invariants in the following. T-invariants correspond to Elementary Modes, which were defined independently and first used in biochemical context [5]. T-invariants and the places in between define minimal functional pathways of the system at steady state. In metabolic systems T-invariants can be interpreted as the minimal amount of enzymes that work at steady state. Thus, the exploration of T-invariants is a crucial step in systems biology.

Unfortunately, the number of T-invariants exponentially grows with network size and complexity. Thus, further decomposition would be very useful. For that reason, we introduced MCT-sets (Maximal Common Transition Sets) [6] and T-clusters [7], which both are based on T-invariants. MCT-sets represent maximal common subsets of the support of T-invariant vectors, where the transitions exclusively occur in the same T-invariants. That means that these transitions work always together, and therefore, the participating enzymes must exhibit a similar expression behavior. The set of transitions of MCT-sets and the places in between with the corresponding edges represent disjunctive subnetworks, which can be interpreted biologically as functional building blocks. Another possibility is to cluster the support of T-invariant vectors using the Tanimoto index as distance measure and applying hierarchical clustering methods such as UPGMA and Neighbor Joining. The resulting set of transitions called T-clusters

and the places in between with the corresponding edges represent overlapping subnetworks, which can be interpreted biologically as functional modules.

These functional building blocks and functional modules can be used for network decomposition validation, exploration, and in synthetic biology.

Whereas Elementary Modes were used mainly for analyzing metabolic networks, T-invariants have been used also to explore signal transduction systems [4, 6] and gene regulatory networks [8].

Results:

In the talk, two examples are explored, one for metabolic networks to illustrate the main idea of T-invariant analysis and the other for a medical application – a model, which combines signal transduction and gene regulation.

The first example is a stoichiometry-based model which describes the main carbon metabolism in potato tubers [9]. The model was built on own experimental data and an existing kinetic model. The qualitative Petri net model has been used to verify the kinetic model and to explore the overall systemŠs behavior. A hierarchical modeling technique was used to represent reversible reactions into one visible vertex. After computation of several network properties, for example checking for deadlocks, the T-invariant analysis was performed. The twelve T-invariants representing functional sub-pathways were explored by hand and illustrate the biological behavior within this metabolism. In this way it was possible to correct and extend the kinetic model.

The second example is related to the Duchenne Muscular Dystrophy (DMD). DMD is one of the most frequently inherited neuromuscular diseases in children. It is an X-linked recessive disease with a birth prevalence of 1 in 3500 live born males. The disease is caused by mutation(s) in the dystrophin gene that result in a loss of the protein dystrophin. This loss is followed by the primary structural dysfunction and in addition it also influences several downstream processes. An efficient therapy is not available. Experimental data suggest signal transduction pathways downstream dystrophin that could compensate the dystrophin defect partially. In order to get new insights and, thus, new ideas for therapeutic possibilities mathematical modeling has been incorporated into research.

This contribution describes modeling and analysis of the first theoretical model downstream the dystrophin gene connecting two main signal transduction pathways encompassing dystrophin and gene regulation of participating proteins, such as transcription factors and utrophin A. The model [8] has been developed on the basis of own experimental data, mainly Real-Time PCR data. Model validation applies invariant analysis, using MCT-sets [6], T-clusters [7], and Mauritius maps [7].

The 107 T-invariants were further decomposed into 25 MCT-sets and 34 T-clusters and explored for their biological meaning. To determine important network parts an exhaustive single-knockout analysis was done. To illustrate the knockout analysis, the corresponding Mauritius Map was derived. Mauritius Maps are represented by a special data structure (binary tree) which describes dependencies between T-invariants.

Analyses of the model resulted in experimental modulation of selected members of the network using human skeletal muscle cells in cell culture whose consequences were studied on mRNA and protein level. The experiments show surprising results, which led to a new iteration of model extension and analysis.

References

[1] Koch and Heiner, in Biological Network Analysis, Eds. B.H. Junker, F. Schreiber, Wiley Book Series in Bioinformatics, chapter 7:139-180 (2008)

[2] Koch, Reisig, Schreiber, Modeling in Systems Biology – The Petri Net Approach, Springer, New York, Berlin (2010)

[3] Doi et al., In Silico Biology 6, 0001, (2006)

[4] Sackmann et al., Computational Biology and Chemistry 31:1-10 (2007)

[5] Schuster et a.l., Proc. Sec. Gauss Symposium 1966:101-114 (1993)

[6] Sackmann et al., BMC Bioinformatics 7:482 (2006)

[7] Grafahrend-Belau et al., BMC Bioinformatics 9:90 (2008)

[8] Grunwald et al., BioSystems 92: 189-205 (2008)

[9] Koch et al., Bioinformatics 21(7):1219-1226 (2005)

Keywords: Systems biology, Petri nets, invariant analysis, T-invariants, MCT-sets, T-clusters, Duchenne Muscular Dystrophy

Joint work of: Koch, Ina; Sackmann, Andrea; Grafahrend-Belau, Eva; Schreiber, Ralf; Ackermann, Jörg; Junker, Björn

Estimation of alternative splicing isoform frequencies from RNA-Seq data

Ion Mandoiu (University of Connecticut - Storrs, US)

We present a novel expectation-maximization algorithm for inference of alternative splicing isoform frequencies from high-throughput transcriptome sequencing (RNA-Seq) data. Our algorithm exploits largely ignored disambiguation information provided by the distribution of insert sizes generated during sequencing library preparation, and takes advantage of base quality scores, strand and read pairing information if available. Empirical experiments on synthetic datasets show that the algorithm significantly outperforms existing methods of isoform and gene expression level estimation from RNA-Seq data.

Keywords: RNA-Seq, alternative splicing isoforms, expectation maximization

Joint work of: Nicolae, Marius (University of Connecticut); Mangul, Serghei (Georgia State University); Mandoiu, Ion (University of Connecticut); and Zelikovsky, Alex (Georgia State University)

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2010/2687

Tree Assignments and their Application in Bioimaging

Axel Mosig (Shanghai Institutes for Biological Sciences, CN)

Generalizing the concept of bipartite matchings to trees leads to the so-called tree-assignment problem. As will be demonstrated in this seminar, tree assignments are highly useful in bioimaging; when combined with certain hierarchical image representations, they allow to compute co-segmentations between two or more images. This yields a novel approach to applications such as cell tracking or spectral image classification. Applied to cell tracking in three dimensions, it allows for reliable cell tracking in cellular environments where established thresholding approaches fail, for instance involving inhomogeneous background or interconnected cells.

Keywords: Bioimaging, co-segmentations, cell tracking, spectral image classification, tree-assignment problem

Joint work of: Mosig, Axel; Hang, Xiao

Random Graphs and Population Genomics

Laxmi Parida (IBM TJ Watson Research Center, US)

The modeling of the evolutionary dynamics of evolving populations as random graphs, if not surprising, is new. The exploration of this began as a quest for understanding the reconstructability of common evolutionary history of populations. This provided interesting insights including a purely topological (or graph theoretic definition) of traditional population genomic entity like the GMRCA (Grand Most Common Ancestor) of individuals or units under mutations as well as recombinations. Apart from giving interesting characterizations of another important structure called the ARG (Ancestral Recombinations Graph), it provided the basis for adapting very naturally the coalescence theory in designing ARG sampling algorithms.

This connection opens the door for many interesting questions that can be posed in the area both from a practical (such as actual detection of recombination events in a population sample) as well as a theoretical standpoint.

Keywords: Population genomics, random graph, coalescent theory, recombiantions

Filter based local multiple alignment tool

Nadia Pisanti (University of Pisa, IT)

I introduce some preliminary results of a new tool we are developing. The tool is a local multiple aligner for biological sequences that makes use of a filter named Tuiuiu as a preprocessing and also uses the information gathered by the filter to speed up the alignment itself. To this purpose I make an overview of the state of the art of filtering methods and tools.

This is a work in progress.

Keywords: Repetitions, biological sequences, local alignment

Joint work of: Pisanti, Nadia; Sagot, Marie-France; Peterlongo, Pierre; Federico, Maria

Efficient computation of statistics for words with mismatches

Cinzia Pizzi (University of Padova, IT)

Since early stages of bioinformatics, substrings played a crucial role in the search and discovery of significant biological signals. Despite the advent of a huge variety of different approaches and models to accomplish these tasks, substrings of fixed length continue to be widely used to determine statistical distributions and compositions of biological sequences at various levels of details. However, in real experimental settings the length m of such patterns is usually restricted to small values, since their exaustive enumeration would lead to a size of the search space that increases exponentially with m. The situation get even worse when some variability is taken into account, and therefore when mismatches are introduced, they are limited to a very small number. Here we overview fast algorithms for computing the counted and expected frequency for words with k mismatches, when it is assumed that the words of interest occurs at least once exactly in the sequence under analysis. Under these settings it is possible to compute the frequency with mismatches of a word that occurs in the input string x of length n in amortized linear time in the input length, independently of the number of mismatches. The expected frequency with mismatches under IID can be computed either in $O(k^2)$ time for any word in the text after $O(kn)$ preprocessing, or in $O(kn)$ time for all the words of some fixed length. For Markov distributions the complexity increases to $O(k|\Sigma|^2)$ after $O(nk|\Sigma|^{p+1})$ preprocessing.

Keywords: Statistics on words, mismatches, dynamic programming, biological sequences

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2010/2738

A fast and accurate approach to detect novel sequences in a de novo human genome assembly

Benjarath Pupacdi (CRI - Bangkok, TH)

Whole human genome sequencing has been advancing at a very rapid rate. The task itself will become routine in a near future. Recently, Li et al. reports an exciting result that the de novo assembly of the 1st Asian genome, the YanHuang genome, contains approximately 5Mbps unique from the reference genome. This makes it imperative to develop efficient methods for detecting novel sequences in a de novo genome assembly, which typically consists of a large number of scaffolds and contigs, e.g. there are nearly 200 thousand scaffolds and contigs of various lengths in the de novo YanHuang genome assembly. In addition, their respective chromosomal positions are unknown, which makes the task of aligning them to a reference genome computationally challenging. In this work, we introduce Novel Sequence Identification Tool (NSIT), which effectively and accurately aligns a de novo assembly against a reference genome and identify its novel sequences.

Keywords: De novo human genome assembly, comparative genomics

Joint work of: Pupacdi, Benjarath; Javed, Asif (IBM TJ Watson Research Center); Zaki, Mohammed J. (Rensselaer Polytechnic Institute)

An Anatomist's tool for exploring 100-dimensional molecular image data

Peter Serocka (Shanghai Institutes for Biological Sciences, CN)

Traditional immunofluorescence microscopy techniques recently have been extended to colocalize, i.e. to simultaneously label and image, a large number of proteins. The Toponome Imaging system (MELC/TIS) by Walter Schubert et. al. (Nat Biotechnol. 2006 Oct;24(10):1270-8) makes is possible to colocalize up to 100 and more proteins in a single fixed probe, aiming at gaining new insights into complex protein-protein functional networks. Various spectroscopy imaging methods are evolving into the same direction, like mass spectroscopy (MALDI) imaging, Fourier-transform Infrared (FTIR) and Rahman spectroscopy imaging.

The challenge for analyzing the resulting multivariate image data sets from all these technologies is: To make both – the contained histological information as well as the high-dimensional protein abundance data – readable to biological and medical experts in an interlinked and yet comprehensive way. This is prerequisite to choosing and applying more automated tools and algorithms that otherwise lack an underlying biological model or guidance.

We describe our visualization software tool LASAGNE which implements two interactive approaches in real time: (1) segmentation by molecular contents and (2) anatomical slicing. Together these methods allow a) for gaining

a fast overview of the high-dimensional data, b) for an in-depth quantitative annotation of histological features on sub-cellular and on supra-cellular scales simultaneously, as well as c) for the comparison of complex data sets.

Data sets shown include MELC/TIS images of human tissue sections affected by the Psoriasis skin disease, and we also demonstrate the application of the Lasagne tool to other imaging techniques like MALDI IMS. New insights gained with LASAGNE are of high relevance by themselves, and we also motivate how LASAGNE can be used to define ground truths and gold standards for further high-throughput molecular imaging experiments.

Keywords: Toponomics, histology, imaging, data mining, visualization, multivariate image data

Joint work of: Serocka, Peter; Schubert, Walter (University Magdeburg); Dress, Andreas (PICB, Shanghai Institutes for Biological Sciences)

Beyond the position weight matrix; and thoughts on uncovering regulatory networks in silico

Rahul Siddharthan (The Institute of Mathematical Sciences - Chennai, IN)

Transcriptional regulation has been a topic of interest for decades, but prediction of transcription factor (TF) binding sites remains largely simplistic and based on the "position weight matrix" (PWM) model. Recently [1] I described a generalisation of the PWM to take account of positional correlations within binding sites, and also observed that there appears to be a sequence signature that extends beyond the "core" binding motifs. Site prediction with PWMs generally yields too many "hits" to be useful, and while including correlations improves the specificity to known functional sites, no purely sequence-analysis-based approach is likely to be perfect. Several studies suggest that TF binding is abundant in the genome and most of this binding may not be functional. Additionally, interspecies comparisons suggest that there is considerable "turnover" of binding sites. So predicting gene regulation should involve not only improved binding site prediction for individual, but prior information of various other kinds: known biochemical roles of the TFs and putative targets, cooperation and competition between different TFs, high-throughput TF binding data (ChIP-chip, ChIP-seq), gene expression data (microarray), information from related organisms, and so on.

This is made more complicated by the fact that much of this information is noisy and unreliable. Integrating uncertain, prior information into the analysis of data is the domain of Bayesian statistics, but there are challenges in implementing this. While I (like many others) am working on a subset of these issues, that I sketch out, I look forward to speculation, discussion, and, hopefully, concrete ideas to emerge at this meeting.

Keywords: Gene regulation, transcription factors, evolution, position weight matrices

Full Paper:
http://www.plosone.org/article/info:doi/10.1371/journal.pone.0009722

See also: R. Siddharthan, PLoS ONE 5(3): e9722 (2010).

Splits and cut sets of tight spans

Andreas Spillner (Universität Greifswald, DE)

With every finite metric space one can associate a polytopal complex known as its tight span. The tight span can help our understanding of certain properties of a given metric space by phrasing these properties in the language of polytopes.

This, for example, might render them more accessible to techniques such as linear programming.

Motivated by applications in phylogenetics which aim to identify relevant partitions of a set of taxa, we recently studied a special type of cut sets of the tight span. These cut sets naturally induce partitions of the ground set of the given metric space. In our work on characterizing when a partition arises from such a cut set, two split indices came up, which, somewhat surprisingly to us, are related in an intriguing way.

Keywords: Metric space, tight span, split index, cut set

Joint work of: Spillner, Andreas; Dress, Andreas (PICB, Shanghai, China); Koolen, Jacobus (POSTECH, South Korea); Moulton, Vincent (University of East Anglia, Norwich, United Kingdom); Wu, Taoyang (PICB, Shanghai, China)

Bioinformatics for Computational Historical Linguistics

Peter F. Stadler (Universität Leipzig, DE)

The evolution of human languages follows principles not unlike those of sequence evolution in molecular biology. It is not surprising therefore that computational approaches developed in molecular phylogenetics can be transferred to linguistic problems. At the same time, there are substantial differences that distinguish sequence evolution from the evolution of words. Maybe most importantly, the words are short so that the establishing homology – identifying cognates – becomes a crucial problem. Historical linguistics, furthermore, focusses on the detailed reconstruction of ancestral states. The alignment problem, for instances, is complicated by the need to infer systematic sound changes from the same data.

We constructed a computational pipeline that implements the typical workflow of a historical linguist with a bioinformatician's toolkit. Applications to two very different language families, Tsezic (spoken in the Caucasus) and Mataco-Guaicuruan (South America) show that this works with acceptable accuracy.

Keywords: Historical linguistics, word alignment

Joint work of: Steiner, Lydia; Stadler, Peter F.; Cysouw, Michael

Computing the Genomic Distance in Linear Time

Jens Stoye (Universität Bielefeld, DE)

The genomic distance problem in the Hannenhalli-Pevzner (HP) theory is the following: Given two genomes whose chromosomes are linear, calculate the minimum number of translocations, fusions, fissions and inversions that transform one genome into the other. We will present a new distance formula based on a simple tree structure that captures all the delicate features of this problem in a unifying way, and a linear-time algorithm for computing this distance.

Keywords: Comparative genomics, genomic distance computation, HP theory.

Joint work of: Bergeron, Anne (Département d'informatique, Université du Québec à Montréal, Canada); Mixtacki, Julia (Universitat Bielefeld, Technische Fakultat, AG Genominformatik, Germany); Stoye, Jens

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2010/2689

Full Paper:
http://dx.doi.org/10.1016/j.tcs.2009.09.008

See also: Theor. Comput. Sci. 410(51), 5300-5316, 2009.

Finding gene regulatory structures in DNA

Esko Ukkonen (University of Helsinki, FI)

The expression of individual genes in mammals is controlled in combinatorial fashion by so-called transcription factors (TFs). Such regulation takes place via binding of a cluster of TFs to a suitable pattern of binding sites of the TFs that is located in the DNA upstream of the gene to be regulated. Such patterns of binding sites are called cis regulatory modules. To predict such modules we have developed a comparative genomics based approach that finds conserved patterns of binding sites by aligning the binding sites in the DNA of two species. The best local alignments under a specific scoring function, found using a Smith-Waterman type algorithm, constitute our predicted regulatory modules. We were able to predict some modules whose regulatory function was then successfully verified in vivo. Moreover, we found an SNP that is associated with increased risk of colorectal cancer. Currently we are working to develop refined models of regulatory models that could be learned from novel large datasets produced by the SELEX procedure.

Keywords: Gene regulation, cis regulatory module, comparative genomics, transcription factor binding site, Smith-Waterman algorithm

Joint work of: Ukkonen, Esko (University of Helsinki); Palin, Kimmo (Sanger Institute); Kivioja, Teemu (Karolinska Institutet); Rastas, Pasi (University of Helsinki); Jolma, Arttu (Karolinska Institutet); Aaltonen, Lauri (University of Helsinki); Taipale, Jussi (Karolinska Institutet)

See also: (1) O. Hallikas, K. Palin, N. Sinjushina, R. Rautiainen, J. Partanen, E. Ukkonen, J. Taipale. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affnity. Cell 124 (January 13, 2006), 47–59.
(2) K. Palin, J. Taipale, E. Ukkonen. Locating potential enhancer elements by comparative genomics using the EEL software. Nature Protocols 1 (2006), 368–374.
(3) S. Tuupanen, M. Turunen, R. Lehtonen, O. Hallikas, S. Vanharanta, T. Kivioja, M. Björklund, Gonghong Wei, Jian Yan, I. Niittymäki, J.P. Mecklin, H. Järvinen, A. Ristimäki, M. Di Bernardo, P. East, L. Carvajal-Carmona, R. S. Houlston, I. Tomlinson, K. Palin, E. Ukkonen, A. Karhu, J. Taipale, L. A. Aaltonen. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nature Genetics 41, 3 (Aug 2009), 885–890.
(4) A. Jolma, L. Cheng, J. Toivonen, T. Kivioja, M. Taipale, J.M. Vaquerizas, J. Yan, M. Sillanpää, M. Bonke, K. Palin, S. Talukder, T.R. Hughes, N.M. Luscombe, E. Ukkonen, J. Taipale. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Research, in press.

Inferring Evolutionary History with Network Models in Population Genomics: Challenges and Progress

Yufeng Wu (University of Connecticut - Storrs, US)

Recently, various network models have been proposed and studied for different evolutionary processes. In population genomics, ancestral recombination graph has been used to model the evolutionary history of population sequences where recombination occurs. In phylogenetics, reticulate networks are used to model reticulate evolutionary processes, such as horizontal gene transfer and hybrid speciation. Although network models may be more accurate in these circumstances, working with network models is challenging computationally.

In this talk, I introduce several computational problems related to the network evolutionary models. The focus is on network models for population genomics. The key biological process is meiotic recombination. I describe several problem formulations related to recombination and the current status of these problems. I also briefly describe my recent work on reticulate networks.

Keywords: Population genomics, reticulate evolution, recombination, algorithms, combinatorial optimization, phylogenetics.

Full Paper:
http://www.engr.uconn.edu/∼ywu

See also: Y. Wu. Close Lower and Upper Bounds for the Minimum Reticulate Network of Multiple Phylogenetic Trees. In Proceedings of ISMB 2010.